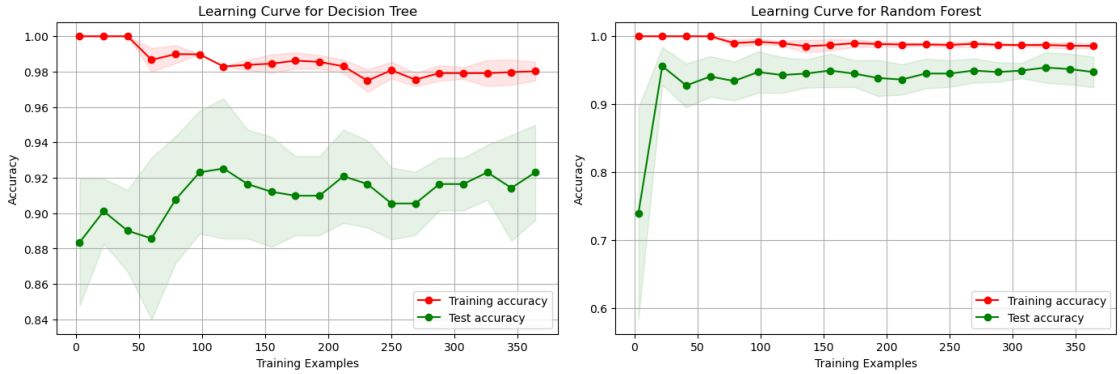
Report

이재우

1. Ensemble Method

1.1.



Decision Tree	Training Accuracy: 0.9758, Test Accuracy: 0.9386
Random Forest	Training Accuracy: 0.9780, Test Accuracy: 0.9494

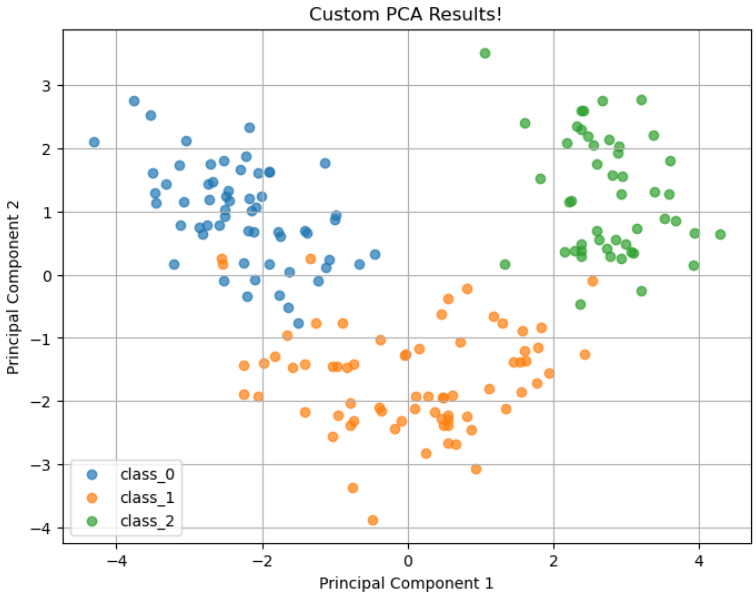
1.2.

두 번째 모델이 테스트 정확도(Accuracy)가 더 높으므로 유방암을 예측하는 모델에 더 적합하다.

Decision Tree 기반 모델은 Training data 에서 accuracy 은 높은 반면, Validation data 의 accuracy 는 0.92 – 0.93 에서 정체되는데 이는 모델이 학습 데이터에 과적합되었음을 암시한다. 반면 Random Forest 기반 모델은 학습 데이터와 검증 데이터에서 일관적으로 매우 높은 수준의 accuracy 를 보여준다.

2. PCA

2.1.



2.2.

PCA 는 주성분을 기준으로 데이터를 압축함으로써 노이즈를 최소화하고 데이터의 품질을 극대화할 수 있다는 장점이 있다. 반면 PCA 는 선형적인 변환을 기반으로 하기 때문에 비선형적인 데이터에 대해서는 적합하지 않다는 한계가 존재한다.

PCA 외에 LDA(Linear Discriminant Analysis), t-SNE(t-Distributed Stochastic Neighbor Embedding) 등의 방법이 차원 축소에 사용된다. LDA 는 PCA 와 유사하나 데이터의 분산이 아닌 클래스의 차이를 최대화하는 축을 찾는다는 특징이 있으며, t-SNE 는 비선형적인 데이터의 시각화에 사용할 수 있는 차원축소기법이다.

3. SVM

3.1.

SVM 은 분류 문제에 적용할 수 잇는 방법론으로서 클래스 사이에 존재하는 margin 을 최대화하는 경계를 찾는 방식이다. Hard margin SVM 은 예외를 허용하지 않는 초평면을 형성하는 방식이다. 이는 노이즈가 없는 데이터에서 분리 성능이 매우 뛰어나지만, 엄격히 구분되지 않는 대부분의 분류 문제에 대해서 적용하기 어렵다는 한계가 존재한다. 반면 Soft margin SVM 은 초평면에 의해 데이터의 클래스가 완벽히 구분되지 않을 수 있다는 전제에 기반하고 있으며, 일부 오류를 허용하되 페널티를 부여하는 방식으로 초평면을 찾아가는 방식을 채택한다. 계산 복잡도가 높으며 하이퍼 파라미터(C)의 적절한 값을 찾기 위한 작업이 필요하다는 한계가 있지만, 노이즈가 있는 데이터에 적용 가능하다는 장점이 있다.

3.2.